

Development of distance formulation for high-dimensional data visualization in multidimensional scaling

Paska Marto Hasugian¹, Herman Mawengkang², Poltak Sihombing³, Syahril Efendi³

¹Doctoral Program, Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

²Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sumatera Utara, Medan, Indonesia

³Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Received May 25, 2024

Revised Oct 3, 2024

Accepted Oct 17, 2024

Keywords:

Distance

Multidimensional scaling

Pasca-multidimensional scaling

Performance

Visualization

ABSTRACT

This research aims to produce a new method called pasca-multidimensional scaling (pasca-MDS) by modifying the multidimensional scaling (MDS) method, the developed model comes as a solution to overcome the problem of data complexity by reducing its description dimension without losing important information. This model, offers an innovative approach in dealing with these problems. Pasca-MDS not only focuses on reducing the dimensionality of data, but also retains the essence of relevant information from each data point. As such, it allows for easier and more efficient analysis without compromising the accuracy of the information conveyed. The main advantage of pasca-MDS lies in its ability to produce simpler visual representations while maintaining the original structure of complex data. This provides clarity and ease in understanding the patterns or relationships hidden within. By using adjustment techniques after the MDS process, this model can provide more optimized results. This process allows the adjustment of data points to achieve a better representation in a lower dimensional space, resulting in a more intuitive and easy-to-understand interpretation. The developed distance formula has the ability to minimize stress compared to other distance formulas in MDS space, with the aim of improving the accuracy of high-dimensional data visualization.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Paska Marto Hasugian

Doctoral Program, Department of Information Technology

Faculty of Computer Science and Information Technology, Universitas Sumatera Utara

Medan, Indonesia

Email: paskamartohasugian@students.usu.ac.id

1. INTRODUCTION

Visualization refers to the process of creating a visual representation of data or information. The goal of data visualization is to communicate information through a visual medium, which allows users to access insights into the data properties of a given data set generally determine which type of visual encoding is more effective [1], [2]. Data visualization plays an important role in revealing relationships and trends that may not be apparent when looking at raw multidimensional data sets [3]. Complex data sets often consist of many interrelated variables, under these conditions, data visualization uses mathematical techniques to reduce the number of dimensions of the data, making it possible to see and understand the relationships between variables in a more intuitive and effective way [4]. Data visualization rests on the premise that a picture is worth a thousand words with the assertion that visualizations have the ability to communicate

information with high intensity and effectiveness, which can replace or surpass lengthy text explanations [5]. Visualization has gradually been placed at the forefront of research in this century and has developed rapidly in the past decade [6]. Data visualization is useful because it transforms complex information into visual forms that are easy to understand, reveals hidden information, facilitates communication, and supports better decision-making in big data [5]. Visualization of high-dimensional data is difficult due to various factors such as number of dimensions, samples, dataset size, density, sparsity, cluster density, and data structure. All these parameters pose a challenge in visualizing complex data [7].

The problem of visualizing high-dimensional data with low sample sizes is often difficult and faces problems such as overfitting, curse of dimensionality, computational infeasibility, and strong model assumptions [8]. Visualization depends on the number of dimensions and the number of samples in the dataset. The higher the number of dimensions, the more complex the data, while a large number of samples can affect the complexity and total size of high-dimensional data [9]. High-dimensional data has challenges such as data management issues, this condition is an increase in the number of features or dimensions can make data modeling and analysis more difficult and complex [10]. In addition, this data continues to grow and change over time [11].

The literature is full of statements that emphasize the importance of visualization. Data visualization transforms complex information into easily understandable visual forms, brings out hidden information, facilitates effective communication, and supports accurate decision-making, and team collaboration in data analysis [11]. Data visualization plays an important role in revealing relationships and trends that may not be apparent when looking at multidimensional data sets. The right visualization approach allows users from various backgrounds, both industrial and academic, to gain valuable insights without having to perform complex calculations [12]. Visualization has a central role in the development of Metaverse as it not only influences its visual construction process, but also determines how users interact and understand the world created [13]. With the significant increase in the amount of data from rapidly developing bioinformatics analysis technologies, the importance of genomic data visualization has become increasingly crucial in facilitating the understanding and efficient analysis of structural variation [14]. Visualization is crucial in explaining the complexity of information. Change-of-use graphs, at-sea distribution maps, and sensor, and layout infographics provide an overview of the evolution of the technology and the area under observation [15]. Visualization has significant benefits in aiding the understanding of the structure and distribution of solutions generated by algorithms, enabling holistic evaluation of performance simplifying interaction, allowing users to select suitable solutions and explore the solution space efficiently and reducing problem complexity [16]. The development of high-dimensional data visualization has become a significant source of innovation and identified promising future research directions. The paper emphasizes dimensionality reduction as a key technique in analyzing and visualizing high-dimensional data [17], [18].

Based on research that has developed data visualization Peterfreund and Gavish [19] identified that ambient noise level is a limitation in data visualization, Hagele *et al.* [20] has developed unsupervised multidimensional scaling (MDS) that considers the uncertainty of data assuming normal distribution and the complexity of calculations becomes a challenge in data visualization. Research by Zhang *et al.* [21] stating that the dependence on iteration which causes high computation time and delivered by Dzemyda *et al.* [22] that the problem is in the interpretation of visualization, so that the development of data dimension reduction will be carried out by ensuring normal data distribution through a series of data transformation processes using Z-score, skewness, kurtosis, and scaling. In addition, this study also developed a distance formulation for MDS which can solve the problem of dependence on iteration which causes high computation time and solve the limitations of visualization interpretation.

2. METHOD

This section introduces the basic model that became the reference in the development of the method, namely MDS, which was later developed into a new model called pasca-multidimensional scaling (pasca-MDS). This model is designed to overcome the limitations of MDS in projecting high-dimensional data to lower dimensions, so as to produce more optimal and accurate visualizations.

2.1. Pasca-multidimensional scaling model architecture

The development of pasca-MDS models requires an in-depth analysis of each process involved to improve the effectiveness of data placement. This process begins with dataset preparation, which includes the use of various data transformation techniques to ensure data quality and consistency. These transformation techniques include skewness analysis to measure the slope of the data distribution, kurtosis to assess the spikiness of the distribution, and scaling to normalize the data so that all variables are on the same scale. After the dataset preparation, the next step is the establishment of the distance formula. This process involves analyzing the established distances, and evaluating the suitability, and accuracy of the existing distance

formulas. Based on this analysis, a new distance formula called distance pas (D_Pas) was developed. The distance pas (D_Pas) formula aims to provide a more accurate and relevant representation of the relationship between data. Overall, the details and steps in this process are comprehensively described in Figure 1.

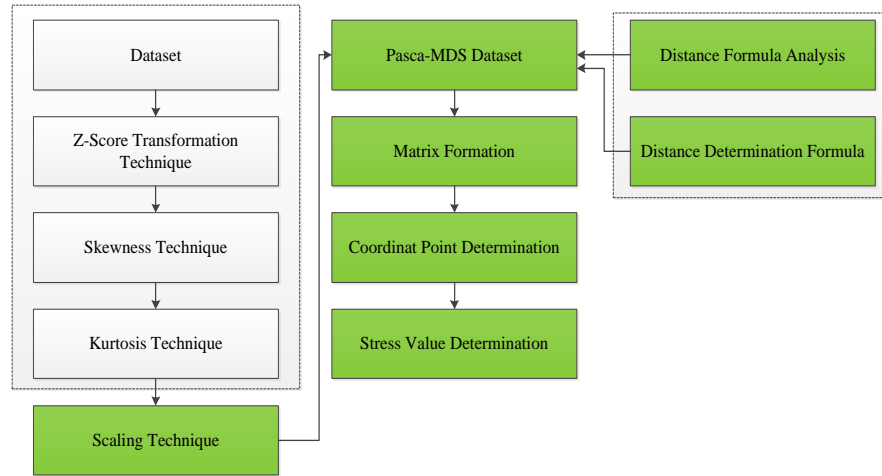


Figure 1. Pasca-MDS model development architecture

2.2. Pasca-multidimensional scaling formulation

The pasca-MDS method developed in this study is designed to project high-dimensional data into lower dimensions, thus facilitating visualization and further analysis. The process involves several systematic steps, from data selection and processing to evaluation and validation of the projected results. Each step plays an important role in ensuring that the resulting data remains accurate and representative of the original structure. The following is a full description of the steps applied in this pasca-MDS method.

a. Z matrix formation

– Matrix formation Z-score normalization:

$$z = \frac{(x_i - \mu)}{\sigma} \quad (1)$$

skewness adjustment using normalized results:

$$\lambda = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \quad (2)$$

the formula will produce a skewness value and test the following conditions in the skewness principle.

$$z_s = \begin{cases} \sqrt[3]{z}, & \text{if skewness} > 0 \\ z^2, & \text{if skewness} < 0 \\ z, & \text{if not} \end{cases}$$

– Formation of kurtosis so that the data transformation is close to normal then the formula is used:

$$k = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (3)$$

provided that,

$$z_{sk} = \begin{cases} \log(1 + z_s), & \text{if kurtosis}(z_s) > 3 \\ e^{z_s} - 1, & \text{if kurtosis}(z_s) < 3 \\ z_s, & \text{otherwise} \end{cases}$$

– Scaled value mapping:

$$X' = \mu + \sigma \cdot z_{sk} \quad (4)$$

- b. Distance formation with formula d_{pas} :

$$d_{pas}(v_i, v_j) = \frac{1}{r} \sqrt{\sum_{k=1}^r \left(\frac{v_{ik} - v_{jk}}{\max(|v_{ik}|, |v_{jk}|)} \right)^2} \quad (5)$$

- c. Find the eigenvalue and eigenvector with the formula $\det(BI)$ and $\det(BI)X$ where to calculate the matrix B with elements:

$$b_{ij} = -\frac{1}{2} (d_{ij}^2 - d_i^2 - d_j^2 - d_{\dots}^2) \quad (6)$$

$$d_i^2 = \frac{1}{n} \sum_i d_{ij}^2$$

$$d_j^2 = \frac{1}{n} \sum_i d_{ij}^2$$

$$d_{\dots}^2 = \frac{1}{n^2} \sum_i d_{ij}^2$$

- d. Form object coordinates based on eigenvectors $X=[X1 \quad X2]$ then calculate D which is the Euclidean distance of the formed coordinates calculate the stress value with (7):

$$S = \left[\frac{\sum_{i=j}^n (d_{ij} - d_{ij})^2}{\sum_{i=j}^n d_{ij}^2} \right] \quad (7)$$

2.3. Distance matrix formulas

As a basis for developing distance formulas, an evaluation of the theory and an evaluation of the distance formulas that have been developed by giving initials to each distance formula, namely D1, D2, D3, D4, D5, D6, D7, D8, D9, and D10 with a description in Table 1.

Table 1. Distance matrix formula

Distance matrix	Formula (v_i, v_j)	Description
Arccosine (D1)	$\text{arccos} \left(\frac{\sum_{k=1}^r v_{ik} v_{jk}}{\sqrt{\sum_{k=1}^r v_{ik}^2} \sqrt{\sum_{k=1}^r v_{jk}^2}} \right)$	Arccosine matrix is used to calculate distance based on the across value of the dot product between two vectors normalized by their magnitude [23].
Canberra (D2)	$\sum_{k=1}^R \frac{ v_{ik} - v_{jk} }{ v_{ik} + v_{jk} }$	The Canberra distance has the task of measuring the distance between two points in a high-dimensional space. Canberra distance is commonly used in data analysis to measure the difference between two feature vectors. This metric is suitable when the data has a large range of values and outliers are present [24].
Dice (D3)	$\frac{\sum_{k=1}^r (v_{ik} - v_{jk})^2}{\sum_{k=1}^r v_{ik}^2 + \sum_{k=1}^r v_{jk}^2}$	The dice distance metric is a metric used to measure the similarity between two samples or feature vectors (v_i) and (v_j). Its function is to measure similarity or distance in the context of common elements shared by two sets or two vectors in data analysis [23].
Divergence (D4)	$2 \sum_{k=1}^R \frac{(v_{ik} - v_{jk})^2}{(v_{ik} + v_{jk})^2}$	Divergence is a measure of the spread or difference between two vectors [23].
Euclidean distance (D5)	$\sqrt{\sum_{k=1}^r (v_{ik} - v_{jk})^2}$	Euclidean distance measurement between two vectors v_i and v_j , where r is the number of dimensions of the vectors. Euclidean distance is often used in various applications such as machine learning, statistics, and data analysis to measure the distance or difference between two points in a multidimensional space [25], [26].
Jaccard (D6)	$\frac{\sum_{k=1}^r (v_{ik} - v_{jk})^2}{\sum_{k=1}^r v_{ik}^2 + \sum_{k=1}^r v_{jk}^2 - \sum_{k=1}^r v_{ik} v_{jk}}$	Jaccard distance is used to measure the similarity and diversity between two data sets or vectors. The formula you mentioned seems to be a variation of the traditional Jaccard index, specialized for vector data and may be more suitable for certain applications such as in data processing or similarity analysis [23].
Lorentzian (D7)	$\sum_{k=1}^r \log(1 + v_{ik} - v_{jk})$	Lorentzian distance formula uses the logarithm of one plus the absolute difference of the components to calculate the distance [26].
Manhattan (D8)	$\sum_{k=1}^r v_{ik} - v_{jk} $	Calculates distance as the sum of the absolute differences between the components of a vector, similar to walking in a city with grid-shaped street [23], [25].
Sorenson (D9)	$\frac{\sum_{k=1}^r v_{ik} - v_{jk} }{\sum_{k=1}^r v_{ik} + v_{jk} }$	Similar to dice, it measures the similarity between two objects [23].

3. RESULTS AND DISCUSSION

Pasca-MDS performance in the formation of data visualization with evaluation based on how well the original data is represented in a lower dimensional space, while changes in the mean of each variable and its standard deviation can indicate differences in data structure after analysis. The normalization graph gives an idea of how well the normalization has been done, while the evaluation of the Euclidean distance after transformation can reveal how well the transformation has been done. Matrix determination can be evaluated by comparing the resulting matrix with the original distance matrix, while the establishment of object coordinates can be evaluated by comparing the object positions in the low-dimensional space with the object relationships in the original space. Finally, the stress value is a measure of the conformity of the data representation in the low-dimensional space with the original data in the high-dimensional space. The architecture in stress value formation in Figure 2.

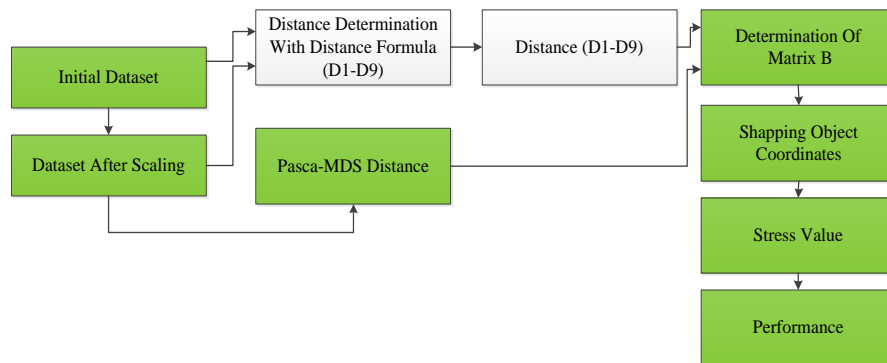


Figure 2. Pasca-MDS performance building architecture

3.1. Performance of pasca-multidimensional scaling data normalization

Establishing normalization is an important step in preparing data sets for visualization. Normalization plays an important role in transforming the data to have a uniform scale, thus facilitating interpretation and analysis. The stages of work to be done include dataset preparation, feature selection, normalization, dataset visualization, evaluation, storage. Pasca-MDS normalization formation architecture in Figure 3.

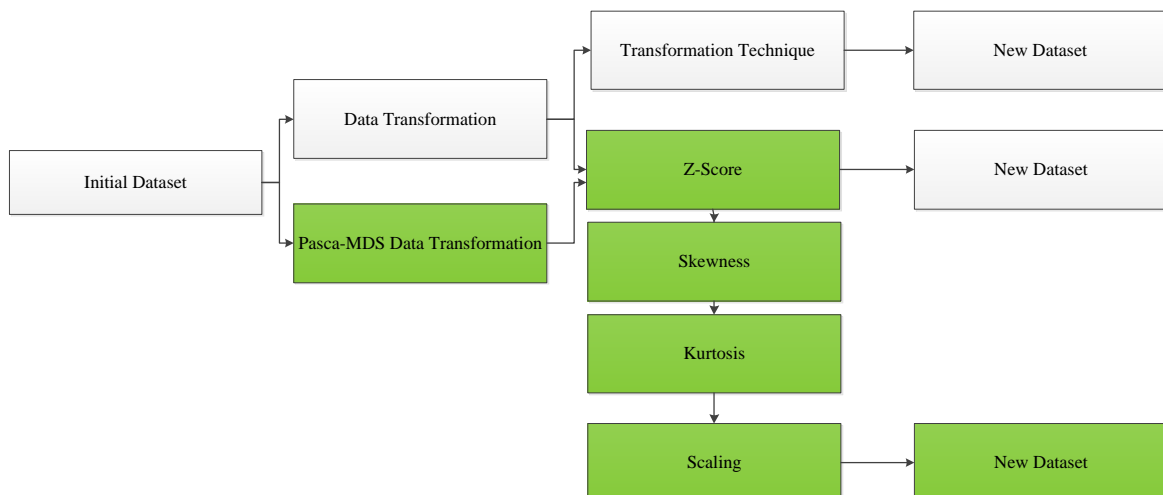


Figure 3. Formation of pasca-MDS data transformation

The objective of normalization after scaling is to produce a more normal distribution of data, where the data distribution tends to be symmetrical and centered around the mean value. This more normal

distribution of data can improve the quality of data visualization by facilitating more effective data interpretation and analysis. The formation of normalization by adjusting the normalization process of Z-score, skewness, kurtosis, and scaling successively produces a histogram in Figure 4 with data distribution towards normal distribution. Figure 4(a) shows the formation of the data distribution using the Z-score, which results in a distribution with a mean around zero and a standard deviation of one, allowing a general understanding of the spread of the data. Furthermore, the skewness analysis in Figure 4(b) reveals asymmetry in the distribution, indicating a potential skewing of the data in one particular direction, while the kurtosis analysis in Figure 4(c) highlights the thickness or height of the tails of the distribution, which may indicate the presence of outliers or extreme variation. After going through the scaling process in Figure 4(d), the data distribution becomes closer to a normal distribution, with a more centered and more even range of values. The last stage is the initial dataset that will be used during the pasca-MDS implementation.

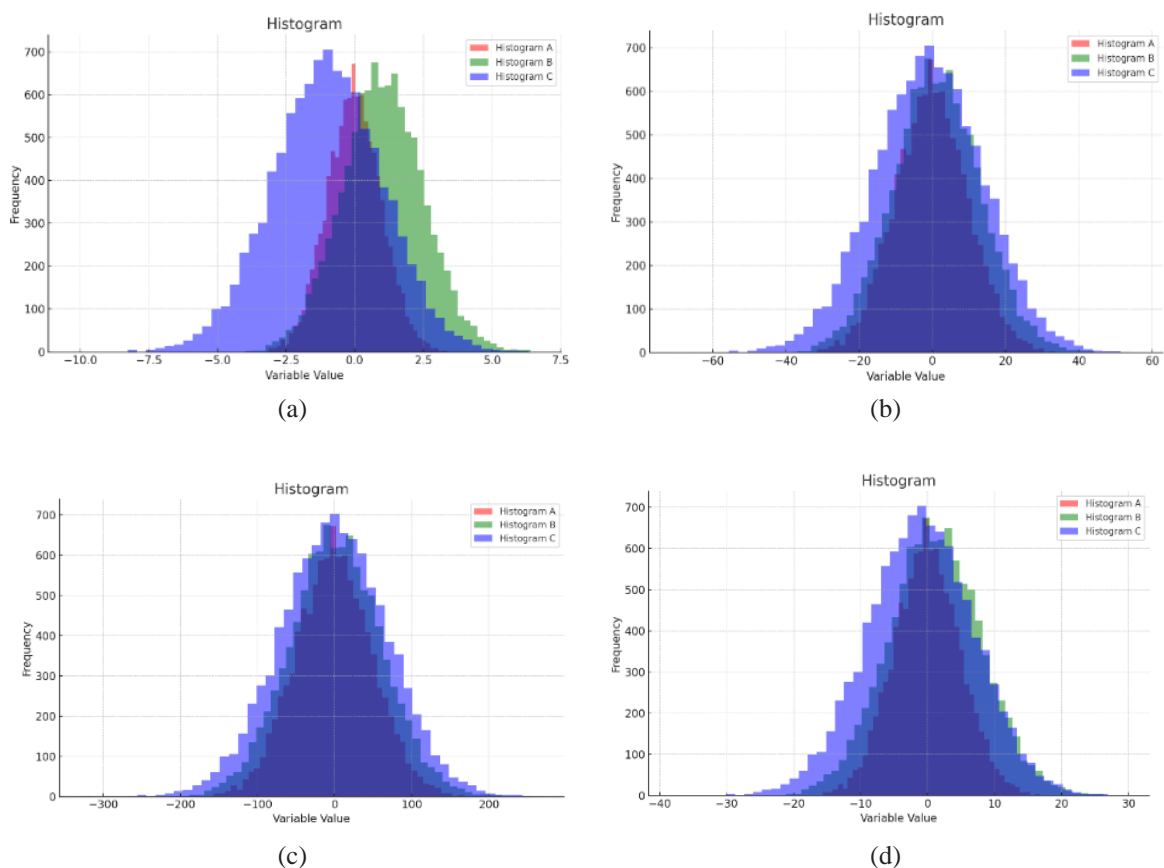


Figure 4. Transformation with: (a) Z-score, (b) skewness, (c) kurtosis, and (d) scaling with normal distribution

3.2. Distance performances

Distance in pasca-MDS is a formula that has been developed in this research, with a specific visualization of distance performance with the architecture of the distance determination process described in Figure 5. The pasca-MDS distance formula, which has been developed in this research, plays an important role in the analysis of multidimensional data structures. This formula is designed to measure the distance between objects in pasca-MDS which is a technique to reduce the dimensionality of data and visualize the relationship between objects in a lower space.

This pasca-MDS distance formula has the advantage of producing a better representation of the original data structure, especially when the data has complex multidimensional properties. Thus, this formula makes an important contribution in understanding and interpreting complex data more effectively. In this research, we will explain in detail about the pasca-MDS distance formula developed, as well as how it can be used to measure the distance between objects more accurately in the context of multidimensional analysis. The formula comparison that forms the basis for the development of the distance formula in pasca-MDS

consists of arccosine distance (D1), canberra distance (D2), dice distance (D3), divergence distance (D4), euclidean distance (D5), jaccard distance (D6), lorentzian distance (D7), manhattan distance (D8), sørenson distance (D9), and pasca-MDS distance (D10). Each type of distance has advantages and disadvantages in determining the distance depending on the data preparation used. In this context, pasca-MDS shows the best distance after normalization with Z-score, normalization after skewness, kurtosis, and scalasis, which are steps in the pasca-MDS process. Graphical analysis shows that the pasca-MDS distance undergoes changes in the determination of distances that are at the same coordinates, and proves to be superior to several other distance formulas. Evaluation is done by utilizing the pasca-MDS distance formula and other distance formulas against the normalization technique with a description of the results. After testing using the distance formula, Table 2 is outlined. Which is a comparison of distance based on the type of matrix against data transformation.

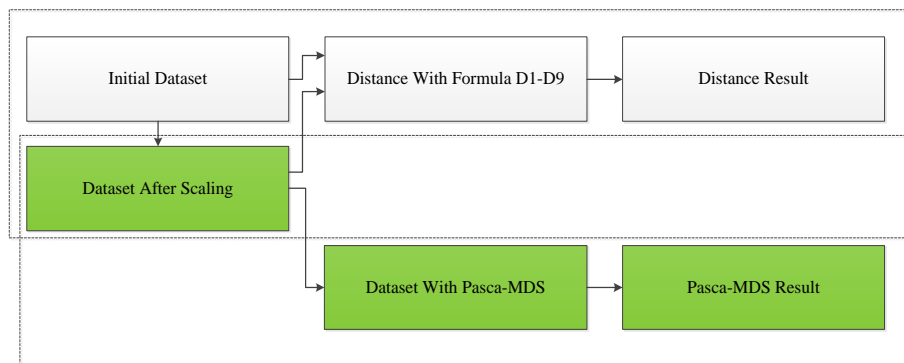


Figure 5. Pasca-MDS distance establishment architecture

Table 2. Distance comparison based on matrix type

Matrix distance/ normalization	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Z-Score	-0.774	0.70406	1.70310	36.02884	2.350225	1.2601	1.3123	2.457394	1.44857	0.551415
Skewness	-0.774	0.70406	1.70310	36.02884	13.86409	1.26010	3.1970	14.496296	1.44857	0.551415
Kurtosis	-0.774	0.70406	1.70310	36.02884	88.53625	1.26010	6.1294	92.573468	1.44857	0.551415
Skalasi	-0.6452	0.64490	1.55449	19.79684	11.31057	1.21706	2.8671	11.731907	1.11896	0.511301

The change in distance determination shown by pasca-MDS distance demonstrates the ability to overcome the challenge of measuring the distance between complex data, thus providing more valuable information in more in-depth data analysis. The smaller the distance value between two data points that are actually related or similar, the better the formula is considered in describing the closeness or similarity between data. This is supported by the MDS work steps in determining the stress value, where a lower stress value indicates a better match between the calculated multidimensional distance and the original distance in the data space. Thus, D10 was found to be more optimal. The graphical visualization with the comparison of each distance against the transformation technique is described in Figure 6 with a comparative study of distances through the transformation of: Z-score (Figure 6(a)), skewness (Figure 6(b)), kurtosis (Figure 6(c)), and scaling (Figure 6(d)).

Based on the evaluation conducted on various distance metrics before and after applying preprocessing techniques, such as Z-score, skewness, kurtosis, and scaling. Before preprocessing, the best shortest distance metric is D3 with a distance value of 0.037959. However, after applying Z-score, skewness, and kurtosis, the shortest distance metric changes to D3 with a distance value of 1.703100. Likewise, after applying scaling, the shortest distance metric changes to D10 with a distance value of 0.511301 and is consistent for all data conditions as shown by the comparison of all distance formulas in Figure 6. This is an assessment of the effectiveness of data transformation in pasca-MDS distance optimization. The data transformations evaluated include Z-score, skewness, kurtosis, and scaling. The evaluation results show that the scaling transformation provides more optimal pasca-MDS distance improvement compared to the other transformations. This illustrates the important role of scaling in improving the multidimensional representation of data in the context of pasca-MDS distance. Therefore, strategies to improve the quality of

multidimensional data analysis can be focused on improving the transformation at the scaling stage. The results of the comparison of distance values between metrics before and after preprocessing are outlined in Figure 7. Figure 7 shows 10 distance values D1-D10 plotted based on four different normalization methods of Z-score, skewness, kurtosis, and scaling. The vertical axis represents the distance value, which shows how far each data point is from the reference point after applying the corresponding normalization. The horizontal axis, labeled as d represents the distance formula. Formula D10 is a developed formula that shows the best distance with consistent coverage resulting in values close to the value of 0.

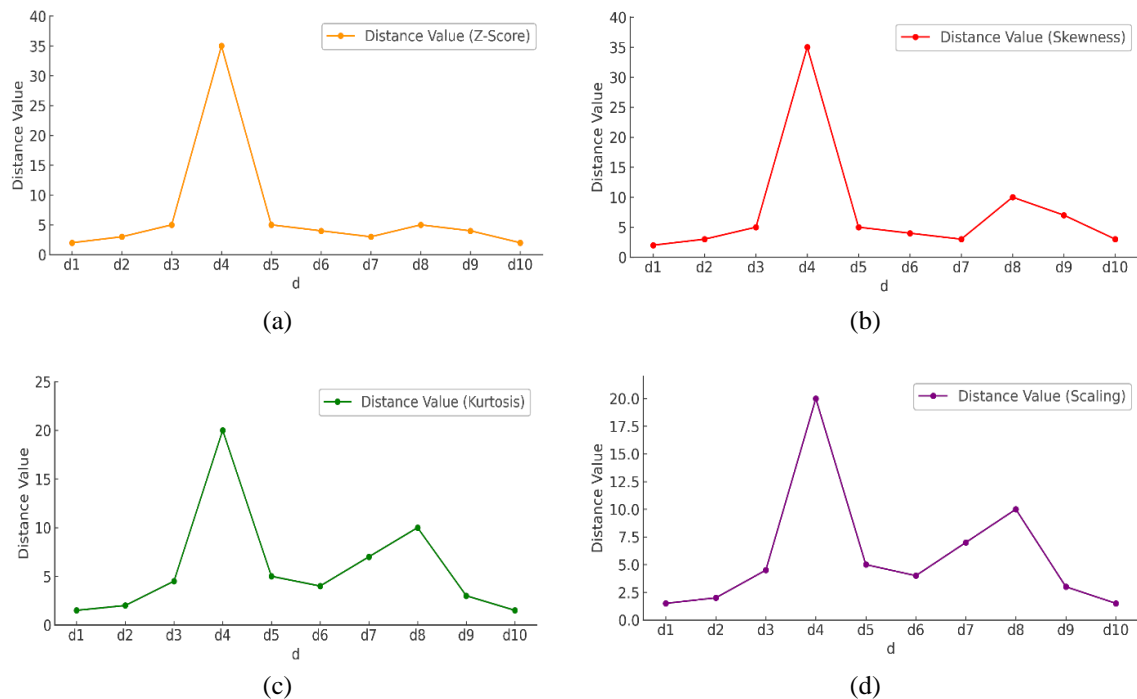


Figure 6. Comparison of 10 distances against normalization: (a) Z-score, (b) skewness, (c) kurtosis, and (d) scaling

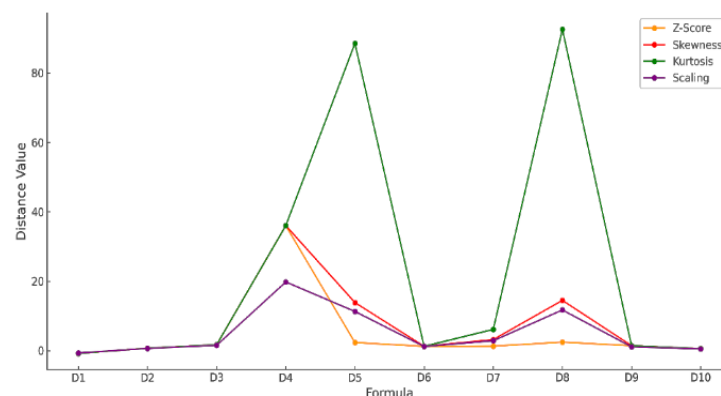


Figure 7. Comparison of distance to the whole transformation process

3.3. Evaluation of pasca-multidimensional scaling stress value

The stress value is a formula used to ensure that the data coordinates are in the right position with the architecture of determining the stress value before and after the use of pasca-MDS. outlined in the following Figure 8. The evaluation of stress values involved periodic data experiments with varying amounts of data, namely 200, 500, 1000, and 2000, conducted with various distance formulas tested using the Python application, the data runs are outlined in Table 3.

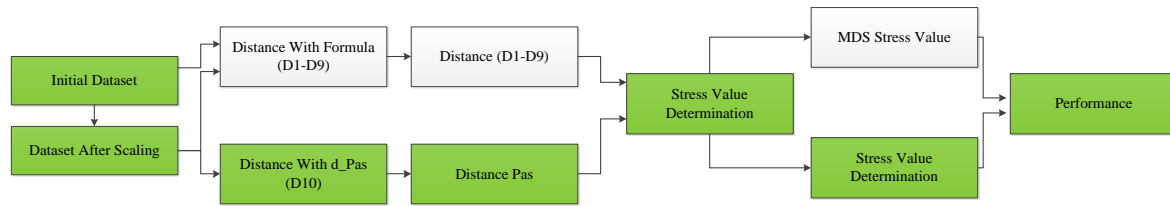


Figure 8. Pasca-MDS stress value evaluation architecture

Table 3. Pasca-MDS stress value performance

Dataset	Stress value
200	0.8447
500	0.9030
1000	0.9311
2000	0.9487

After generating the stress values using pasca-MDS, the comparison results with the distance formula are outlined on tables and graphs to provide a better understanding of the performance of the distance formula on different datasets. This will help in determining the most appropriate distance formula for data of a particular size, as well as provide an understanding of how the size of the dataset affects the resulting stress values. the representation of the distance between objects is declared optimal if the value of S is close to 0, which means that it will be at the optimal point if the value is close to the value of 0, which indicates that the representation of the distance in the resulting data (d_{ij}) is very close to the actual distance in the data (d_{ij}). The comparison results are described in Table 4.

Table 4. Comparison of stress values

Matrix formula	Stress values			
	Dataset 200	Dataset 500	Dataset 1000	Dataset 2000
D1	0.9201	0.9496	0.9645	0.9728
D2	0.9364	0.9603	0.9718	0.9790
D3	0.8776	0.9232	0.9459	0.9583
D4	0.9981	0.9998	0.9998	0.9995
D5	0.9811	0.9888	0.9935	0.9954
D6	0.8593	0.9109	0.9367	0.9526
D7	0.9649	0.9786	0.9932	0.9951
D8	0.9873	0.9873	0.9957	0.9969
D9	0.9441	0.9641	0.9757	0.9808
D_pas	0.8447	0.9030	0.9311	0.9487

Table 4 is a representation of stress values for several datasets of different sizes (200, 500, 1000, and 2000) using matrix formulas (D1-D9 and D_pas). Stress values are used to measure the quality of data embedding in multidimensional analysis, such as in factor analysis or multidimensional mapping. Each cell in the table shows the stress value generated by a particular matrix formula for the corresponding dataset. Lower stress values indicate better data embedding (less information lost in the multidimensional representation), while high stress values indicate the opposite. From Table 4, it can be seen that the larger the dataset size (from 200 to 2000), generally the stress values are lower for all matrix formulas, indicating an improvement in data embedding quality with larger dataset sizes. The best distance order for all phases with a dataset of 200. In this phase, the matrix formula D_pas has the lowest stress value, which is 0.8447. Therefore, the best order for phase 200 is D_pas, D6, D3, D1, D5, D7, D9, D2, D8, and D4. In this phase, the D_pas matrix formula also has the lowest stress value, which is 0.9030. Thus, the best order for phase 500 is D_pas, D6, D3, D1, D5, D7, D9, D2, D8, and D4. Dataset of 1000. In this phase, the D_pas matrix formula has the lowest stress value, which is 0.9311. Thus, the best order for phase 1000 is D_pas, D6, D3, D1, D5, D7, D9, D2, D8, and D4. Dataset of 2000. In this phase, the matrix formula D_pas also has the lowest stress value, which is 0.9487. Therefore, the best order for phase 2000 is D_pas, D6, D3, D1, D5, D7, D9, D2, D8, and D4. The comparison for each stress value is described through the graph in Figure 9.

Figure 9 provides a detailed visualization of the relationship between the various formulas and the values of the stress value for four different dataset sizes, namely 200, 500, 1000, and 2000. The vertical axis

(y-axis) of this graph displays the stress value values that ranges from 0 to 1. This value indicates how much pressure or experienced by the system or object as a result of the application of a particular formula, where values close to zero are used. Formula, where values closer to zero indicate better performance in reducing the stress or pressure experienced. In other words, the lower the stress value value, the better the formula is at maintaining system stability. stability of the system. Meanwhile, the horizontal axis (x-axis) displays the various formulas tested, denoted as D1 to D_pas. These formulas are applied to datasets of different sizes to test how each formula affects the stress value value. The graph it also uses different colors to represent the dataset size, blue for dataset 200, orange for dataset 500, green for dataset 1000, and red for dataset 2000. Based on the analysis results of the visualization, it can be concluded that the D_Pas matrix formula provides the best data embedding quality compared to other matrix formulas. This is shown by the lowest stress value value of D_Pas for all tested dataset sizes. Therefore, D_Pas can be considered as the most effective and reliable distance metric in reducing stress, making it the best choice in this study. The consistent performance of D_Pas, regardless of the dataset size, confirms its superiority as an optimal formula for maintaining stability and accuracy in the data embedding process.

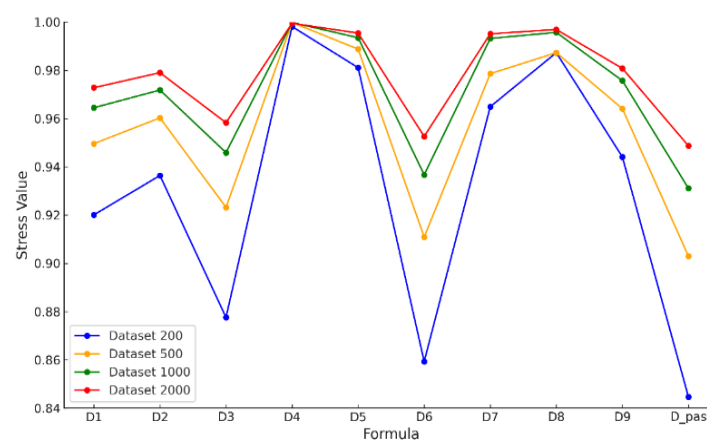


Figure 9. Comparison of pasca-MDS stress values

4. CONCLUSION

Pasca-MDS, as an extension of MDS, offers an innovative approach by retaining relevant information from each data point. The evaluation results show the superiority of the pasca-MDS method with lower stress values with other distance formulas in terms of the nine distances made for comparison in this study. The implication of the research is the need for the development of more efficient data visualization techniques in handling the complexity of high-dimensional data, thus providing an important basis for future research and innovation of better data visualization techniques with a focus on accuracy.

ACKNOWLEDGMENTS

The authors would like to express their deep appreciation to the Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, and the supervisors for their support and guidance throughout the research process. Thanks are also due to the anonymous reviewers for their valuable and constructive feedback, which has helped improve the quality of this research. The author hopes that this work can make a meaningful contribution to the development of science and become a useful foundation for further research, as well as support studies in related fields to enrich the insights of future researchers.




REFERENCES

- [1] S. Elnawawi, L. C. Siang, D. L. O'Connor, and R. B. Gopaluni, "Interactive visualization for diagnosis of industrial Model Predictive Controllers with steady-state optimizers," *Control Engineering Practice*, vol. 121, Apr. 2022, doi: 10.1016/j.conengprac.2021.105056.
- [2] G. Bergk, B. Shariati, P. Safari, and J. K. Fischer, "ML-assisted QoT estimation: A dataset collection and data visualization for dataset quality evaluation," *Journal of Optical Communications and Networking*, vol. 14, no. 3, pp. 43–55, Mar. 2022, doi: 10.1364/JOCN.442733.
- [3] H. Chung, S. Nandhakumar, and S. Yang, "GridSet: Visualizing Individual Elements and Attributes for Analysis of Set-Typed Data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 2983–2998, Aug. 2022, doi: 10.1109/TVCG.2022.3150000.




- 10.1109/TVCG.2020.3047111.
- [4] A. Boaro *et al.*, “Visualization, navigation, augmentation. The ever-changing perspective of the neurosurgeon,” *Brain and Spine*, vol. 2, Elsevier BV, pp. 1-13, 2022, doi: 10.1016/j.bas.2022.100926.
 - [5] L. M. Poste and C. F. Patterson, “Multidimensional Scaling – Sensory Analysis of Yoghurt,” *Canadian Institute of Food Science and Technology Journal*, vol. 21, no. 3, pp. 271–278, 1988, doi: 10.1016/S0315-5463(88)70817-2.
 - [6] T. Jiang, Y. Hou, and J. Yang, “Literature Review on the Development of Visualization Studies (2012–2022),” *Engineering Proceedings*, 2023, vol. 38, no. 1, p. 89, doi: 10.3390/engproc2023038089.
 - [7] M. C. Schatz *et al.*, “Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space,” *Cell Genomics*, vol. 2, no. 1, Jan. 12, 2022, doi: 10.1016/j.xgen.2021.100085.
 - [8] K. Li, F. Wang, L. Yang, and R. Liu, “Deep feature screening: Feature selection for ultra high-dimensional data via deep neural networks,” *Neurocomputing*, vol. 538, p. 126186, Jun. 2023, doi: 10.1016/j.neucom.2023.03.047.
 - [9] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.
 - [10] E. Dimara, H. Zhang, M. Tory, and S. Franconeri, “The Unmet Data Visualization Needs of Decision Makers Within Organizations,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 12, pp. 4101–4112, Dec. 2022, doi: 10.1109/TVCG.2021.3074023.
 - [11] E. Sherman and L. G. Schiffman, “Quality-of-life (QOL) assessment of older consumers: A retrospective review,” *Journal of Business and Psychology*, vol. 6, no. 1, pp. 107–119, 1991, doi: 10.1007/BF01013687.
 - [12] L. M. Ränger, M. von Kurnatowski, M. Bortz, and T. Grützner, “Multi-Objective Optimization of Dividing Wall Columns and Visualization of the High-Dimensional Results,” *Computers & Chemical Engineering*, vol. 142, Nov. 2020, doi: 10.1016/j.compchemeng.2020.107059.
 - [13] Y. Zhao *et al.*, “Metaverse: Perspectives from graphics, interactions and visualization,” *Visual Informatics*, vol. 6, no. 1, pp. 56–67, Mar. 01, 2022, doi: 10.1016/j.visinf.2022.03.002.
 - [14] Y. Lei *et al.*, “Overview of structural variation calling: Simulation, identification, and visualization,” *Computers in Biology and Medicine*, vol. 145, Jun. 01, 2022, doi: 10.1016/j.compbiomed.2022.105534.
 - [15] Z. Yang *et al.*, “UAV remote sensing applications in marine monitoring: Knowledge visualization and review,” *Science of the Total Environment*, vol. 838, Elsevier B.V., Sep. 10, 2022, doi: 10.1016/j.scitotenv.2022.155939.
 - [16] R. Ding, H. bin Dong, G. sheng Yin, J. Sun, X. dong Yu, and X. bin Feng, “An objective reduction method based on advanced clustering for many-objective optimization problems and its human-computer interaction visualization of pareto front,” *Computers & Electrical Engineering*, vol. 93, Jul. 2021, doi: 10.1016/j.compeleceng.2021.107266.
 - [17] P. M. Hasugian, H. Mawengkang, P. Sihombing, and S. Efendi, “Review of High-Dimensional and Complex Data Visualization,” in *2023 International Conference of Computer Science and Information Technology (ICOSNIKOM)*, 2023, pp. 1–7, doi: 10.1109/ICoSNiKOM60230.2023.10364377.
 - [18] S. Liu, D. Maljovec, B. Wang, P. T. Bremer, and V. Pascucci, “Visualizing High-Dimensional Data: Advances in the Past Decade,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 3, pp. 1249–1268, Mar. 2017, doi: 10.1109/TVCG.2016.2640960.
 - [19] E. Peterfreund and M. Gavish, “Multidimensional scaling of noisy high dimensional data,” *Applied and Computational Harmonic Analysis*, vol. 51, pp. 333–373, 2021, doi: 10.1016/j.acha.2020.11.006.
 - [20] D. Hagele, T. Krake, and D. Weiskopf, “Uncertainty-Aware Multidimensional Scaling,” in *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 23–32, 2023, doi: 10.1109/TVCG.2022.3209420.
 - [21] Z. Zhang, D. Wang, B. Yang, and J. Yin, “Weighted Multidimensional Scaling Localization Method With Bias Reduction Based on TOA,” *IEEE Sensors Journal*, vol. 23, no. 17, pp. 19803–19814, 2023, doi: 10.1109/JSEN.2023.3296986.
 - [22] G. Dzemyda, M. Sabaliauskas, and V. Medvedev, “Geometric MDS Performance for Large Data Dimensionality Reduction and Visualization,” *Informatica*, vol. 33, no. 2, pp. 299–320, 2022, doi: 10.15388/22-INFOR491.
 - [23] A. M. Lopes and J. A. T. Machado, “Multidimensional scaling and visualization of patterns in global large-scale accidents,” *Chaos, Solitons and Fractals*, vol. 157, Apr. 2022, doi: 10.1016/j.chaos.2022.111951.
 - [24] E. Blanco-Mallo, L. Morán-Fernández, B. Remeseiro, and V. Bolón-Canedo, “Do all roads lead to Rome? Studying distance measures in the context of machine learning,” *Pattern Recognition*, vol. 141, 2023, doi: 10.1016/j.patcog.2023.109646.
 - [25] M. Raeisi and A. B. Sesay, “A Distance Metric for Uneven Clusters of Unsupervised K-Means Clustering Algorithm,” *IEEE Access*, vol. 10, pp. 86286–86297, 2022, doi: 10.1109/ACCESS.2022.3198992.
 - [26] B. Olea, “Canonical variation of a Lorentzian metric,” *Journal of Mathematical Analysis and Applications*, vol. 419, no. 1, pp. 156–171, 2014, doi: 10.1016/j.jmaa.2014.04.064.

BIOGRAPHIES OF AUTHORS






Paska Marto Hasugian    is a student in the Computer Science Doctoral Program at the Universitas Sumatera Utara Computer Science, Universitas Sumatera Utara 2022. He is the author who completed his undergraduate education in the Informatics Engineering Study Program, at Budidarma University Medan, completed the Master of Computer Studies Program from UPI YPTK Padang, and is currently completing his Doctoral Studies in Computer Science at the Universitas Sumatera Utara. His research interests are in the field of data visualization and data mining. He can be contacted at email: paskamarto86@gmail.com.






Herman Mawengkang    is a Professor at the Faculty of Mathematics, Universitas Sumatera Utara, Medan, Indonesia. Completed his bachelor of mathematics education at the Faculty of Mathematics, Universitas Sumatera Utara, Medan in 1974. Completed his doctorate program in mathematics education at the University of New South Wales, Sydney, Australia. Current research is conducted in the fields of computer science and applied mathematics. Where he is the author/co-author of more than 153 research publications. He can be contacted at email: hmawengkang@yahoo.com, mawengkang@usu.ac.id



Poltak Sihombing    received his Ph.D. in computer science from Universiti Sains Malaysia, Malaysia, in 2010. He is currently a Professor at the Department of Information Technology, Universitas Sumatera Utara, Medan, Indonesia. Since 2021, he has also been the Head of the Computer Science Doctoral Study Program, Universitas Sumatera Utara. is an author who completed his undergraduate education at the Faculty of Mathematics, Universitas Sumatera Utara, Medan. His expertise and skills are in the areas of microcontrollers, and IoT, intelligent systems, and information retrieval. He can be contacted at email: poltak@usu.ac.id.



Syahril Efendi    is a Professor at the Faculty of Computer Science and Information Technology, Universitas Sumatera Utara. From 2021, he also became Secretary of the Doctoral Study Program of Computer Science Universitas Sumatera Utara. is an author who completed his undergraduate education at the Faculty of Mathematics, Universitas Sumatera Utara, Medan. Completed master of computer education in Computer Science Study Program at Universiti Kebangsaan Malaysia (UKM), and completed doctorate program education at Faculty of Mathematics, Universitas Sumatera Utara, Medan. He can be contacted at email: syahril@usu.ac.id.